DOCUMENT RESUME

ED 272 519 TM 860 424

AUTHOR Forbes, Dean W.
TITLE The Rasch Model as a Practical and Effective Procedure for Educational Measurement.
PUB DATE Apr 86
NOTE 12p.; Paper presented at the Symposium, "Taming the Rasch Tiger: Using Item Response Theory in Practical Educational Measurement," at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 17-20, 1986). For other symposium papers, see TM 860 425-428.
PUB TYPE Viewpoints (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Adaptive Testing; Computer Assisted Testing; Elementary Secondary Education; *Item Banks; *Latent Trait Theory; Mathematical Models; Microcomputers; Standardized Tests; *Test Construction; *Test Items; *Test Theory
IDENTIFIERS Calibration; Item Parameters; *Rasch Model

ABSTRACT

For many years personalization of achievement testing has been impossible in all but the simplest forms. Recently, item response theory (IRT), or latent trait theory, has emerged as a valuable tool which brings far greater flexibility to the process than had previously been possible. The single parameter Rasch Model, a mathematical model developed by George Rasch, has proven particularly useful in a number of instances in which it has been used. These situations range from individual school districts, through consortia of school districts, to governmental agencies. IRT makes possible the measurement of performance in relation to latent traits in a way which is especially appropriate to basic skills achievement testing in the public schools. Recent developments in microcomputer technology and the development of computer programs make it practical for the local school district to develop and calibrate files of items to meet its own testing needs. Calibrated item files are becoming more available for purchase by those potential users who would sooner not develop their own. The transition from item file to finished test is well within the scope of the local user. Available procedures range from the intensively clerical to the highly computerized and are well documented in readily available publications. (JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED272519

THE RASCH MODEL AS A PRACTICAL AND EFFECTIVE PROCEDURE FOR EDUCATIONAL MEASUREMENT

Dean W. Forbes
Portland Public Schools
Portland, Oregon

Paper presented at the annual meeting of the National Council for Measurement in Education

San Francisco, California
April, 1986

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D.W. Forbes

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 860 424

THE RASCH MODEL AS A PRACTICAL AND EFFECTIVE PROCEDURE FOR EDUCATIONAL MEASUREMENT

DEAN W. FORBES
PORTLAND (OREGON) PUBLIC SCHOOLS

INTRODUCTION

Education has always been faced with the problem of measuring achievement status and growth. In such measurement, testing people have always wrestled with the problem of matching each child with an appropriate test. It is all too apparent that problems occur when a child is faced with a test which is far removed from current performance level (either too difficult or easy). In either case there is frustration and, in the case of the overly difficult test, a great deal of pain. This problem of arriving at an optimum fit between a child's capability and test difficulty has been one that for years has plagued the testing field.

SCHOOL ACHIEVEMENT TESTING

There have been many attempts to improve the measurement of student performance both for individuals and groups. One of the great developments of the twentieth century was the establishment of tests for specific grade levels which could be normed on a supposedly national population. Theoretically this permitted comparing any individual child's performance with that of others at the same grade level using "national" norms. (It is not appropriate at this point to enter into any debate about the validity or utility of allegedly national norms.) The grade level test was a valuable improvement. It still fell far short of need, however, since many lower performing children found such a test completely beyond their ability.

In an effort to improve the situation there has been, in more recent years, a decided move toward out-of-level testing (where a low performing fifth grader might get a test originally designed and intended for fourth graders). This strategy created some problems in its own right since it then was difficult to interpret performance in light of norms either for the child's actual grade or the grade for which the test was originally intended. Various vertical equating efforts have attempted to provide a basis for interpreting out-of-level testing in meaningful terms.

If a school system is satisfied with the curriculum content of commercially published tests this might be an acceptable compromise to the testing dilemma. However, some residual problems still remain. Such tests, available from commercial publishers, are of necessity planned to handle all children at a given grade level (many test levels in fact span two grades). This means that the difficulty range of such a test is necessarily rather wide. There will be some very easy items to fit the lower performing student and there still have to be some very difficult items to fit the higher performing student; the average student for a given grade level can be nicely handled by the middle difficulty range of the test. Still, at any performance level there will be many items inappropriate to a given child. These non-functional items serve no useful purpose but do take additional testing time and increase the effort required on the part of the student.

If school districts embark on a program of local test development they will find that they have basically the same problems although they might not be as aware of them. One great improvement that is possible in district test development programs is planning and building tests to fit the specific local curriculum (rather than accept the compromises which a commercial developer must make). However, out-of-level testing will have the same problems as are encountered with commercially developed tests, and this may be further complicated by the fact that reference to any type of "national" norm is quite difficult.

Ideally, it should be possible to present each student with a test that is appropriate to that person's current performance. In other words, the test is of a difficulty level that can be handled with comfort. Furthermore, the test should have a range of difficulty narrow enough so that all (or at least most) of the items would be reasonable tasks for the particular child. If such a test system could be built to fit the local curriculum, the measurement millenium would approach.

In order to achieve this goal one would need a series of many tests (each of which would be short and built in terms of a restricted difficulty range), sequenced in order of difficulty, so that any child could at all times be fitted with a test appropriate to present performance capability. Although this need was recognized for years the test development and norming problems seemed insurmountable.

RECENT DEVELOPMENTS IN TEST THEORY

In recent years new developments have emerged which address these problems. Unfortunately, they have brought with them new vocabulary and concepts which seem to alienate, rather than entice, the very user group which stands to gain the most.

"IRT," the "Rasch model," and "latent trait" are three terms which have recently risen to the status of "buzz words." They are frequently used in social situations whether or not they are really understood by the users. The somewhat better informed dilletante, like the naive wine taster, can simulate profundity by exercising a "smattering of ignorance." This might, for instance, involve reference to the "one parameter" versus "three parameter" controversy. This practice may easily intimidate those practical measurement people who lack extensive technical backgrounds and thus do a great disservice to one of the most dramatic and valuable developments in the recent history of test theory.

Rather than inform people about vital new testing tools, overly technical and jargonish explanations all too often becloud the relevant issues. This can create anxiety and lead to avoidance on the part of the very people who would most profit from the use of these new tools.

There is nothing mystical or esoteric about these (and other closely related) terms or the concepts they represent. Let's find out what they are and how we can use them.

IRT simply means "Item Response Theory." It is a generic term which refers to any of a number of theoretical positions which relate to a mathematical model the probability that a person will correctly answer a test item. Such models occur in varying degrees of complexity. (A thorough exposition of this can be found in Lord and Novick, especially in chapters 17-20 which were authored by Allan Birnbaum (5)).

The Rasch Model is a mathematical model developed by George Rasch, a Danish statistician. It is one of the models included under the general heading of IRT.

A Latent Trait is a stable and consistent characteristic underlying behavior which is, itself, unobservable. Knowledge of a person with respect to a latent trait provides a basis for predicting behavior. Examples of possible latent traits of interest to educators would be reading comprehension or arithmetic computation.

Now that we have defined some terms let's return to the issues and problems in educational measurement and see how we can cope with them. Let's find out how to apply Item Response Theory to their solution. Discussion will focus on the Rasch model itself.

LATENT TRAITS AND THEIR MEASUREMENT

Education has always tried to teach in terms of developmental continua. These have been recognized in the basic tool skills (such as reading, mathematics, and language usage) as well as in many more specialized instructional sequences.

It has always been held that certain stages are prerequisite to later stages. This has long formed a conceptual frame for education. As curriculum diversity emerges (after basic skill levels), each branch of instruction may suggest a possible continuum of its own while at the same time showing a divergence from other curriculum content areas. (Piaget spoke to the same issues in physical and cognitive development.) Thus, a conceptual map of education shows an hierarchical organization both within and between its various elements. Although teaching can be so organized, the learning continua underlying the curriculum have consistently been measured in rather haphazard fashion not only in the classroom, but by the traditional wide range standardized achievement test. In summary, this instructional model has been hampered by many measurement problems.

With acceptance of the premise that a measurement continuum should parallel a learning/teaching continuum, it was recognized that if a learning continuum could be adequately measured by an underlying scale extending through its entire range, a student could enter and exit from the measurement continuum at points appropriate to that student's current development regardless of age or grade level. It would follow that scores for such continuum segments could meaningfully be compared either over time or between students. If this could be done, an extremely flexible educational testing system would evolve. The measurement of such an hypothetical testing model remained an unobtainable ideal for years.

In 1967 Benjamin Wright delivered a landmark paper at the ETS Invitational Conference on Testing Problems (10). This paper described a test development system based on one of the many measurement models developed by George Rasch (of Demark) and his colleagues (8). This measurement model (popularly called The Rasch model), as described by Wright, seemed to promise a solution to many of the problems that we have been discussing since: 1.) "scores" were based on the curriculum difficulty of test items, not on the collective response of a comparison group; and 2) the difficulty of items was measured in terms of the probability that a person of given "ability" would get the item correct, rather than the proportion of persons in some reference group who made the correct response.

This model would permit selecting those items from a parent item bank with difficulty levels that matched the present ability of a student, it would further permit relating a "score" on that particular test to a measurement scale underlying the entire difficulty range represented in the total pool of items in the bank.

Thus, the specific model is based on one item parameter (difficulty) and one person parameter (ability). (Ability as used by Rasch and Wright, and in the present paper, merely means capability within whatever subject domain is being measured. It has absolutely no connotation of intelligence or any similar generic ability.) The model is probabilistic (i.e., it builds on the probability of "this person" getting "this item" right). It requires nothing else in order to arrive at a score describing performance but it does make some assumptions, and rather stringent assumptions too, about the nature of the items and of the educational element being measured.

The Rasch model assumes that all items will have equal (or nearly equal) discrimination and that all items involve minimal guessing. The model also assumes, with respect to the educational element being measured, an underlying latent trait of uni-dimensional nature. In other words what is measured is a continuum, not some composite involving content from two or more discrete or inter-twined continua.

The item assumptions must be met during the processes of item writing and bank development (and through test administration practices). The latent trait must be evaluated by fit of individual items to the model (probably after some preliminary intellectual analysis and soul searching).

Given a collection of items that fit the model a person may be tested with a selection of items appropriate in difficulty to that person's present capability and the raw score can be converted to a scaled score in terms of the scale describing the latent trait. Merely the number of items correct is a sufficient statistic for determining performance level.

This provides a tremendous amount of flexibility since a test could, in theory, be prepared for each individual student. It would also be possible to develop tests for particular performance levels and then administer those tests to all appropriate students. It would even be possible to develop wide range tests which could be used in similar fashion to those which have been a part of education for so many years. This latter procedure may at times be appropriate, although it will carry with it many of the problems discussed

earlier. It would be possible to generate short criterion-referenced tests to determine when a student reaches some particular performance level. In fact, so long as the items fit the model it seems to be a test development "Shmoo" (which, for those of you who do not remember the Al Capp cartoons of many years ago, was a "critter" that was all things and all people -- a truly universal specific to meet all needs.)

In practical test usage this means that any person can be measured in terms of present need, and that no two persons must take the same test. This would lend itself well to tailored testing either in the form of printed booklet or computer presentation. In group situations each student could get a "suit off the rack" (i.e., an appropriate and good fitting test from a pre-established series of tests). (Over the past ten years the Portland, Oregon, Public School District has built an entire achievement testing program based on this model (4, 7).

There are a number of other latent trait models. A three parameter one has received much attention and study. It adds parameters based on item discrimination (slope of the difficulty function of the item) and the level of chance performance (guessing). The three parameter model is, of necessity, more complex than the one parameter model.

In general, the single parameter model offers some very real advantages. In the process of developing banks of appropriate test items and validating them for latent trait measurement (a process called calibration) the three parameter model is complicated by the fact that item calibration requires larger samples of respondents than does the one parameter model (which can satisfactorily calibrate an item with 200 or fewer responses (7, No. 20)). Several computer programs are readily available for the calibration of items for the one parameter model and some of these make modest demands for computing power (16). The one parameter model has proved to be very successful in a wide variety of practical educational situations. It has been used by the public schools in Portland Oregon, other member districts of the Northwest Evaluation Association, school districts in California (particularly San Jose), in England, and in Scotland (3,6,7,9). The basic procedures involved in the calibration of items using the single parameter Rasch model are described in Wright's Invitational Conference Paper (10), and Wright and Panchapakesan (15). The basic procedures for building and using tests from calibrated item banks are described in two research memoranda from the Statistical Laboratory of the Department of Education, University of Chicago, by Wright and Douglas (13, 14). A more complete presentation is given in Wright and Stone (16).

ITEM GENERATION AND ITEM BANK BUILDING

An essential ingredient in employing the Rasch model (or any other IRT model) in test construction is a pool of relevant test items which have been demonstrated to fit the model. This means that the items have been calibrated (a process which gives all of the traditional item analysis information plus information verifying fit of the item to the Rasch Model). Computationally the calibration process is laborious but straight forward. Current microcomputers can handle calibration and appropriate programs are readily available. Among those available are BICAL, which deals with the one

parameter model, and LOGIST, dealing with the three parameter model. (In Best Test Design Wright and Stone outline a procedure for hand calibration of items which is actually practical for small situations (16).)

Assuming the availability of skilled item writers, it is possible for a school district to generate its own items. If the district has access to an appropriate computer, it can also calibrate them. If a school district is not able to do this by itself, it is entirely possible to join forces with other interested school districts to pool resources. The Northwest Evaluation Association is an example of such a consortium. It involves independent school districts, primarily from the States of Oregon and Washington. The association also includes representation from California and various other states.

If one does not wish to embark on an item bank development project (with its necessary item calibration) there is a steadily increasing number of Rasch calibrated item banks in existence (some of which are available for purchase (12)). One well established set of item banks (in reading, mathematics, and language skills) is available for purchase from the Northwest Evaluation Association.

An item bank can occur in more than one form. Each of these has certain implications for use. In the simplest form it can even be a file of item cards without an index (to preserve the user's sanity any bank larger than 50 or 60 items needs an index, and even those that are smaller should have one). Large files of item cards can be managed manually with adequate indexing. Indexing and index referral are greatly facilitated by computerization; this permits a great saving in clerical time by letting the computer identify those items meeting the test constructor's specifications. In this way, all non-relevant items can be ignored. (A strictly clerical application would involve scrutinizing each item card to determine whether or not it met the requirements specified by the test author. It doesn't take much imagination to see that this could become a very time consuming and laborious process.)

A higher stage of item bank development puts the items themselves in computer storage (either on hard or floppy disks). For many years practicality limited computer item storage to textual or numerical material with graphics being stored, by necessity, in a supplementary card file. More recent developments, however, permit computer storage of graphics and the introduction of laser beam technology promises even greater sophistication in the near future.

ITEM BANK USAGE FOR TEST ASSEMBLY

Once a user has access to an item bank, a wide variety of tests can be constructed to meet many measurement needs. The process of planning and assembling a test will follow basically the same steps with which you are all familiar. Goals to be measured must be identified, the number of items per goal must be decided, and the specific items to fit this blueprint must be selected. (Remember, we are dealing with a bank of items already written and calibrated. If the contents of the item bank cover the goals to be tested it will be necessary to generate few, if any, additional items.) Now, however, you can add something new to the process. You decide the level of difficulty which you want the test to measure, and the range of difficulty it should

span. At this point testing is being done "on purpose" instead of "by accident." It is even possible to build a series of tests to the same blueprint but with planned differences in difficulty level. This permits accommodating individual differences in a way previously not possible while still preserving the ability to convert raw scores from different tests to a common scale. (This is the city-wide achievement testing model that has been developed and is in use in the schools of Portland, Oregon.)

Tests, regardless of purpose (i.e., unit or semester examinations, criterion referenced achievement tests, or survey achievement tests), can be assembled using a variety of procedures.

They can be put together by purely clerical means. This would involve selecting appropriate item cards which would then be turned over to a typist for preparation of final print copy.

Item cards can be selected, superimposed in page format and placed in an acetate sheath for photocopying of "print copy" pages (this substantially reduces the possibility of clerical error in item reproduction).

Once items are computer stored it is possible for them to be selected and displayed on the computer screen, or printed out by purely electronic means. This permits a number of possibilities. All items can be listed and displayed for final editorial checks and selection. With adequate computer programming it permits actual computer printing of "print copy" pages. It makes possible (again with adequate computer programming) computer assisted testing, whereby the computer presents the student with one item at a time on the video screen and the student uses the computer keyboard as a response mechanism.

This last procedure offers several alternative procedures with differing degrees of sophistication: 1) computer presentation of one pre-constructed test; 2) computer presentation of one test selected from a group of pre-established tests; 3) a tailored test where the computer "jumps" a person to an appropriate item difficulty range and then gives a pre-selected sequence of items; or 4) a computerized adaptive test where the computer initially establishes an estimated performance level and then individually selects each succeeding item based on the correctness or incorrectness of the previous response. In this case testing would involve as few items as are necessary to document the performance level at a predetermined level of measurement error. At that time testing would be ended.

The Scottish Ministry of Education has subsidized the development of some very promising item banks and test construction procedures which have been built around microcomputers and related equipment (2, 3). These projects are programmed so that the computer selects items to meet blueprint specifications set up by the test author. In selection of items the computer program has available various default procedures that can be used in the event that item substitutions must be made due to incompleteness of the basic item files. The computer then formats and prints the actual test pages (including graphics) so that all that is left is the printing. At that point the teacher (or other user) has a test available which fits the measurement needs of the specific situation and which is composed of high quality items. And all at minimal expenditure of effort.

Computerized item bank and test development facilities of this sort can be developed so that they are strictly "stand alone" operations or they can be developed so that a remote station can interact with a larger central processor (and item banks) by means of telephone circuitry.

SUMMARY

For many years personalization of achievement testing has been impossible in all but the simplest forms. Recently, item response theory has emerged as a valuable tool which brings far greater flexibility to the process than had previously been possible. The single parameter Rasch Model has proven particularly useful in a number of instances in which it has been used. These situations range from individual school districts, through consortia of school districts, to governmental agencies.

IRT makes possible the measurement of performance in relation to latent traits in a way which is especially appropriate to basic skills achievement testing in the public schools.

Recent developments in micro-computer technology and the development of computer programs, make it practical for the local school district to develop and calibrate files of items to meet its own testing needs. Calibrated item files are becoming more available for purchase by those potential users who would sooner not develop their own.

The transition from item file to finished test is well within the scope of the local user. Available procedures range from the intensively clerical to the highly computerized and are well documented in readily available publications.

0062E

REFERENCES

1. Bagnall, G.M., Milne, P., and Pollitt, A.B., Primary Mathematics Item Bank (Explanatory Manual). Scottish Education Department, Godfrey Thompson Unit, Department of Education, University of Edinburgh (1982).

2. Bagnall, G.M., Milne, P., and Pollitt, A.B., Primary Mathematics Item Bank (Summary Report). Scottish Education Department, Godfrey Thompson Unit, Department of Education, University of Edinburgh (1984).

3. Godfrey Thompson Unit, Item Banking Facility (Introduction). Godfrey Thompson Unit, Department of Education, University of Edinburgh (Feb. 1984).

4. Kingsbury, G. Gage, A Comparison of Item Response Theory Procedure for Assessing Response Dimensionality. Paper presented at the annual meeting of NCME, Chicago, Ill., 1985.

5. Lord, Frederic M., and Novick, Melvin R., Statistical Theory of Mental Test Scores. Addison-Wesley Publishing Company, 1968.

6. Martois, John S., Measurement Issues Related to the Use of Item Response Theory Based Life Skills Reading Tests. (Personal communicationfrom Dr. John Davis, San Jose, California, Unified School District.)

7. Occasional Papers, Portland (Oregon) Public Schools Department of Research and Evaluation (various authors, topics, and dates).

8. Rasch, George, An Individualistic Approach to Item Analysis. Readings in Mathematical and Social Science. Lazarsfeld and Henry (eds.), Chicago: SRA Inc., 1966.

9. Rentz, Robert R., and Bashaw, W.L., The National Reference Scale for Reading: An Application of the Rasch Model. Journal of Educational Measurement, V14, No. 2 (Summer 1977).

10. Wright, Benjamin D., Sample Free Test Calibration and Person Measurement. In invitational Conference on Testing Problems, Educational Testing Service, 1967.

11. Wright, Benjamin D., Solving Problems with the Rasch Model. Journal of Educational Measurement, V14, No. 2, (Summer 1977).

12. Wright, Benjamin D. and Bell, Susan R., Item Banks: What, Why, and How. Journal of Educational Measurement, V21, No. 4 (Winter 1984).

13. Wright, Benjamin D., and Douglas, Graham A., Best Test Design and Self-Tailored Testing. Research Memorandum No. 19, (June 1975), Statistical Laboratory, Department of Education, The University of Chicago.

14. Wright, Benjamin D., and Douglas, Graham A., Better Procedures for Sample Free Item Analysis. Research Memorandum No. 20, Statistical Laboratory, Department of Education, The University of Chicago.

15. Wright, Benjamin D., and Panchepakesan, Nargis, A Procedure for Sample Free Item Analysis. Educational and Psychological Measurement, V29, No. 1 (Spring 1969).

16. Wright, Benjamin D., and Stone, Mark H., Best Test Design, MESA Press, Chicago, IL. (1979).

0062E